8 *The process–performance paradox in expert judgment*

How can experts know so much and predict so badly?

COLIN F. CAMERER AND ERIC J. JOHNSON

1. INTRODUCTION

A mysterious fatal disease strikes a large minority of the population. The disease is incurable, but an expensive drug can keep victims alive. Congress decides that the drug should be given to those whose lives can be extended longest, which only a few specialists can predict. The experts work around the clock searching for a cure; allocating the drug is a new chore they would rather avoid.

In research on decision making there are two views about such experts. The views suggest different technologies for modeling experts' decisions so that they can do productive research rather than make predictions. One view, which emerges from behavioral research on decision making, is skeptical about the experts. Data suggest that a wide range of experts like our hypothetical specialists are not much better predictors than less expert physicians, or interns. Furthermore, this view suggests a simple technology for replacing experts – a simple linear regression model (perhaps using medical judgments as inputs). The regression does not mimic the thought process of an expert, but it probably makes *more* accurate predictions than an expert does.

The second view, stemming from research in cognitive science, suggests that expertise is a rare skill that develops only after much instruction, practice, and experience. The cognition of experts is more sophisticated than that of novices; this sophistication is presumed to produce better predictions. This view suggests a model that strives to mimic the decision policies of experts – an "expert (or knowledge-based) system" containing lists of rules experts use in judging longevity. An expert system tries to match, not exceed, the performance of the expert it represents.

In this chapter we describe and integrate these two perspectives. Integration comes from realizing that the behavioral and cognitive science approaches have different goals: Whereas behavioral decision theory emphasizes the *performance* of experts, cognitive science usually emphasizes differences in experts' *processes* (E. Johnson, 1988).

A few caveats are appropriate. Our review is selective; it is meant to emphasize the differences between expert performance and process. The generic

decision-making task we describe usually consists of repeated predictions, based on the same set of observable variables, about a complicated outcome – graduate school success, financial performance, health – that is rather unpredictable. For the sake of brevity, we shall not discuss other important tasks such as probability estimation or revision, inference, categorization, or trade-offs among attributes, costs, and benefits.

The literature we review is indirectly related to the well-known "heuristics and biases" approach (e.g., Kahneman, Slovic, & Tversky, 1982). Our theme is that experts know a lot but predict poorly. Perhaps their knowledge is biased, if it comes from judgment heuristics or they use heuristics in applying it. We can only speculate about this possibility (as we do later, in a few places) until further research draws the connection more clearly.

For our purposes, an expert is a person who is experienced at making predictions in a domain and has some professional or social credentials. The experts described here are no slouches: They are psychologists, doctors, academics, accountants, gamblers, and parole officers who are intelligent, well paid, and often proud. We draw no special distinction between them and extraordinary experts, or experts acclaimed by peers (cf. Shanteau, 1988). We suspect that our general conclusions would apply to more elite populations of experts,[1] but clearly there have been too few studies of these populations.

The chapter is organized as follows: In section 2 we review what we currently know about how well experts perform decision tasks, then in section 3 we review recent work on expert decision processes. Section 4 integrates the views described in sections 2 and 3. Then we examine the implications of this work for decision research and for the study of expertise in general.

2. PERFORMANCE OF EXPERTS

Most of the research in the behavioral decision-making approach to expertise has been organized around performance of experts. A natural measure of expert performance is predictive accuracy; later, we discuss other aspects. Modern research on expert accuracy emanates from Sarbin (1944), who drew an analogy between clinical reasoning and statistical (or "actuarial") judgment. His data, and the influential book by Meehl (1954), established that in many clinical prediction tasks experts were *less* accurate than simple formulas based on observable variables. As Dawes and Corrigan (1974, p. 97) wrote, "the statistical analysis was thought to provide a floor to which the judgment of the experienced clinician could be compared. The floor turned out to be a ceiling."

[1] While presenting a research seminar discussing the application of linear models, Robyn Dawes reported Einhorn's (1972) classic finding that three experts' judgments of Hodgkin's disease severity were uncorrelated with actual severity (measured by how long patients lived). One seminar participant asked Dawes what would happen if a certain famous physician were studied. The questioner was sure that Dr. So-and-so makes accurate judgments. Dawes called Einhorn; the famous doctor turned out to be subject 2.

2.1. A language for quantitative studies of performance

In many studies, linear regression techniques are used to construct statistical models of expert judgments (and to improve those judgments) and distinguish components of judgment accuracy and error.[2] These techniques are worth reviewing briefly because they provide a useful language for discussing accuracy and its components.

A subject's judgment (denoted Y_s) depends on a set of informational cues (denoted $X_1, \ldots, X_n$). The cues could be measured objectively (college grades) or subjectively by experts (evaluating letters of recommendation). The actual environmental outcome (or "criterion") (denoted Y_e) is also assumed to be a function of the same cues.

In the comparisons to be described, several kinds of regressions are commonly used. One such regression, the "actuarial" model, predicts outcomes Y_e based on observable cues X_i. The model naturally separates Y_e into a predictable component $\hat{Y}_e$, a linear combination[3] of cues weighted by regression coefficients $b_{i,e}$, and an unpredictable error component Z_e. That is,

$$\begin{aligned} Y_e &= \sum b_{i,e} X_i + z_e \qquad \text{(actuarial model)} \\ &= Y_e + z_e \end{aligned} \tag{1}$$

Figure 8.1 illustrates these relationships, as well as others that we shall discuss subsequently.

2.2. Experts versus actuarial models

The initial studies compared expert judgments with those of actuarial models. That is, the correlation between the expert judgment Y_s and the outcome Y_e (often denoted r_a, for "achievement") was compared with the correlation between the model's predicted outcome $\hat{Y}_e$ and the actual outcome Y_e (denoted R_e).[4]

Meehl (1954) reviewed about two dozen studies. Cross-validated actuarial models outpredicted clinical judgment (i.e., R_e was greater than r_a) in all but one study. Now there have been about a hundred studies; experts did better in only a handful of them (mostly medical tasks in which well-developed theory outpredicted limited statistical experience; see Dawes, Faust, & Meehl,

[2] Many regression studies use the general "lens model" proposed by Egon Brunswik (1952) and extended by Hammond (1955) and others. The lens model shows the interconnection between two systems: an ecology or environment, and a person making judgments. The notation in the text is mostly lens-model terminology.

[3] Although the functions relating cues to the judgment and the outcome can be of any form, linear relationships are most often used, because they explain judgments and outcomes surprisingly well, even when outcomes are known to be nonlinear functions of the cues (Dawes & Corrigan, 1974).

[4] The correlation between the actuarial-model prediction and the outcome Y_e is the square root of the regression R^2, and is denoted R_e. A more practical measure of actuarial-model accuracy is the "cross-validated" correlation, when regression weights derived on one sample are used to predict a new sample of Y_e values.

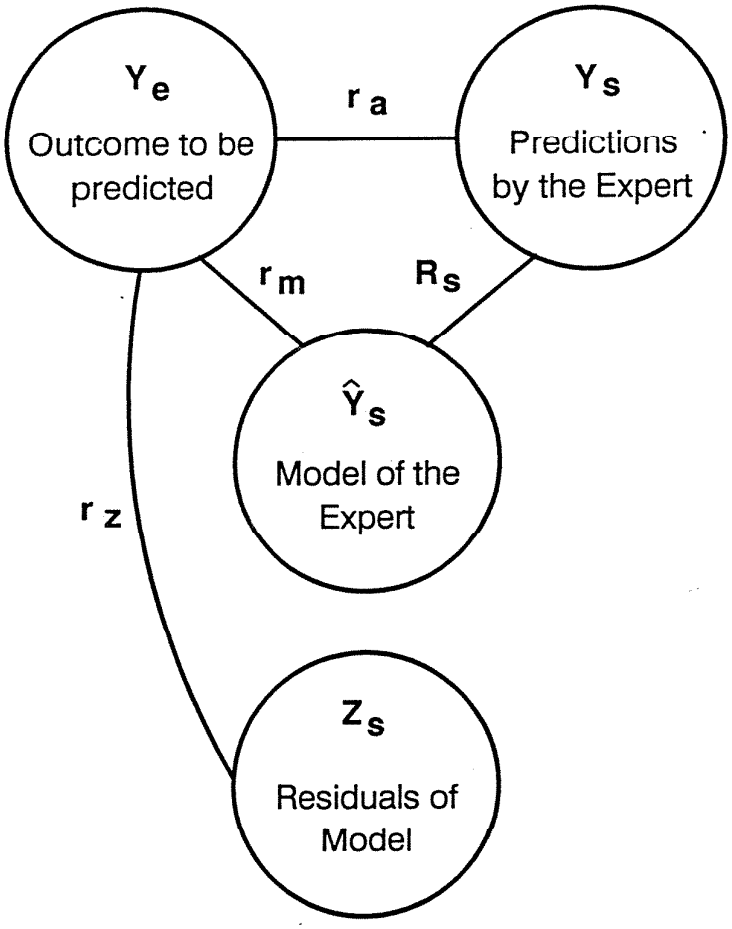

Figure 8.1. A quantitative language for describing decision performance.

1989). The studies have covered many different tasks – university admissions, recidivism or violence of criminals, clinical pathology, medical diagnosis, financial investment, sports, weather forecasting. Thirty years after his book was published, Meehl (1986, p. 373) suggested that "there is no controversy in social science that shows such a large body of qualitatively diverse studies coming out so uniformly in the same direction."

2.3. Experts versus improper models

Despite their superiority to clinical judgment, actuarial models are difficult to use because the outcome Y_e must be measured, to provide the raw data for deriving regression weights. It can be costly or time-consuming to measure outcomes (for recidivism or medical diagnosis), or definitions of outcomes can be ambiguous (What is "success" for a Ph.D.?). And past outcomes must be used to fit cross-validated regression weights to predict current outcomes, which makes models vulnerable to changes in true coeffi-

cients over time. Therefore, "improper"[5] models – which derive regression weights without using Y_e – might be more useful and nearly as accurate as proper actuarial models.

In one improper method, regression weights are derived from the Y_s judgments themselves; then cues are weighted by the derived weights and summed. This procedure amounts to separating the overall expert judgment Y_s into two components, a modeled component $\hat{Y}_s$ and a residual component z_s, and using only the modeled component $\hat{Y}_s$ as a prediction.[6] That is,

$$\begin{aligned} Y_s &= \Sigma b_{is} X_i + z_s \\ &= \hat{Y}_s + z_s \end{aligned} \tag{2}$$

If the discarded residual z_s is mostly random error, the modeled component $\hat{Y}_s$ will correlate more highly with the outcome than will the overall judgment, Y_s. (In standard terminology, the correlation between $\hat{Y}_s$ and Y_e, denoted r_m, will be higher than r_a.)

This method is called "bootstrapping" because it can improve judgments without any outcome information: It pulls experts up by their bootstraps. Bowman (1963) first showed that bootstrapping improved judgments in production scheduling; similar improvements were found by Goldberg (1970) in clinical predictions based on MMPI scores[7] and by Dawes (1971) in graduate admissions. A cross-study comparison showed that bootstrapping works very generally, but usually adds only a small increment to predictive accuracy (Camerer, 1981a). Table 8.1 shows some of those results. Accuracy can be usefully dissected with the lens-model equation, an identity relating several interesting correlations. Einhorn's (1974) version of the equation states

$$r_a = r_m R_s + r_z(1 - R_s^2)^{1/2} \tag{3}$$

where R_s^2 is the bootstrapping model R^2 (how closely the judge resembles the linear model), and r_z is the correlation between bootstrapping-model residuals z_s and outcomes Y_e (the "residual validity"). If the residuals z_s represent only random error in weighing and combining the cues, r_z will be close to zero. In this case, r_m will certainly be larger than r_a, and because $R_s \leq 1$, bootstrapping will improve judgments. But even if r_z is greater than zero (presumably because residuals contain some information that is correlated with outcomes), bootstrapping works unless

[5] By contrast, actuarial models often are called "optimal linear models," because by definition no linear combination of the cues can predict Y_e more accurately.

[6] Of course, such an explanation is "paramorphic" (Hoffman, 1960): It describes judgments in a purely statistical way, *as if* experts were weighing and combining cues in their heads; the process they use might be quite different. However, Einhorn, Kleinmuntz, and Kleinmuntz (1979) argued persuasively that the paramorphic regression approach might capture process indirectly.

[7] Because suggested Minnesota Multiphasic Personality Inventory (MMPI) cutoffs were originally created by statistical analysis, it may seem unsurprising that a statistical model beats a judge who tries to mimic it. But the model combines scores *linearly*, whereas judges typically use various scores in configural nonlinear combinations.

$$r_z \geq r_m \left(\frac{1 - R_s}{1 + R_s} \right)^{1/2} \tag{4}$$

For $R_s = .6$ (a reasonable value; see Table 8.1), residual validity r_z must be about half as large as model accuracy for experts to outperform their own bootstrapping models. This rarely occurs.

When there are not many judgments, compared with the number of variables, the regression weights in a bootstrapping model cannot be estimated reliably. Then one can simply weight the cues equally[8] and add them up. Dawes and Corrigan (1974) showed that equal weights worked remarkably well in several empirical comparisons (the accuracies of some of these are shown in the column r_{ew}, in Table 8.1). Simulations show that equal weighting generally works as well as least squares estimation of weights unless there are twenty times as many observations as predictors (Einhorn & Hogarth, 1975). As Dawes and Corrigan (1974) put it, "the whole trick is to decide what variables to look at and then to know how to add" (p. 105).

2.4. Training and experience: experts versus novices

Studies have shown that expert judgments are less accurate than those of statistical models of varying sophistication. Two other useful comparisons are those between experts and novices and between experienced and inexperienced experts.

Garb (1989) reviewed more than fifty comparisons of judgments by clinical psychologists and novices. The comparisons suggest that (academic) training helps but additional experience does not. Trained clinicians and graduate students were more accurate than novices (typically untrained students, or secretaries) in using the MMPI to judge personality disorders. Students did better and better with each year of graduate training. The effect of training was not large (novices might classify 28% correctly, and experts 40%), but it existed in many studies. Training, however, generally did *not* help in interpreting projective tests (drawings, Rorschach inkblots, and sentence-completion tests); using such tests, clinical psychologists probably are no more accurate than auto mechanics or insurance salesmen.

Training has some effects on accuracy, but experience has almost none. In judging personality and neurophysiological disorders, for example, clinicians do no better than advanced graduate students. Among experts with varying amounts of experience, the correlations between amount of clinical experience and accuracy are roughly zero. Libby and Frederick (1989) found that experience improved the accuracy of auditors' explanations of audit errors only slightly (although even inexperienced auditors were better than students).

In medical judgments too, training helps, but experience does not. Gustaf-

[8] Of course, variables must be standardized by dividing them by their sample standard deviations. Otherwise, a variable with a wide range would account for more than its share of the variation in the equally weighted sum.

Table 8.1. *Examples of regression-study results*

Study	Prediction task	Mean accuracy of: Model fit, R_s	Judge, r_a	Bootstrapping model, r_m	Bootstrapping residuals, r_z	Equal-weight model, r_{ew}	Actuarial model,[a] R_e
Goldberg (1970)	Psychosis vs. neurosis	.77	.28	.31	.07	.34	.45
Dawes (1971)	Ph.D. admissions	.78	.19	.25	.01	.48	.38
Einhorn (1972)	Disease severity	.41	.01	.13	.06	n.a.	.35
Libby (1976)[b]	Bankruptcy	.79	.50	.53	.13	n.a.	.67
Wiggens & Kohen (1971)	Grades	.85	.33	.50	.01	.60	.57

[a]All are cross-validated R_e except Einhorn (1972) and Libby (1976).
[b]Figures cited are recalculations by Goldberg (1976).
Source: Adapted from Camerer (1981a) and Dawes & Corrigan (1974).

son (1963) found no difference between residents and surgeons in predicting the length of hospital stay after surgery. Kundel and LaFollette (1972) reported that novices and first-year medical students were unable to detect lesions from radiographs of abnormal lungs, but fourth-year students (who had had some training in radiography) were as good as full-time radiologists.

These tasks usually have a rather low performance ceiling. Graduate training may provide all the experience one requires to approach the ceiling. But the myth that additional experience helps is persistent. One of the psychology professors who recently revised the MMPI said that "anybody who can count can score it [the MMPI], but it takes expertise to interpret it." (*Philadelphia Inquirer,* 1989). Yet Goldberg's (1970) data suggest that the only expertise required is the ability to add scores with a hand calculator or paper and pencil.

If a small amount of training can make a person as accurate as an experienced clinical psychologist or doctor, as the data imply, then lightly trained paraprofessionals could replace heavily trained experts for many routine kinds of diagnoses. Citing Shortliffe, Buchanan, and Feigenbaum (1979), Garb (1989) suggested that "intelligent high school graduates, selected in large part because of poise and warmth of personality, can provide competent medical care for a limited range of problems when guided by protocols after only 4 to 8 weeks of training."

It is conceivable that outstanding experts are more accurate than models and graduate students in some tasks. For instance, in Goldberg's (1959) study of organic brain damage diagnoses, a well-known expert (who worked very slowly) was right 83% of the time, whereas other Ph.D. clinical psychologists got 65% right. Whether such extraordinary expertise is a reliable phenomenon or a statistical fluke is a matter for further research.

2.5. Expert calibration

Whereas experts may predict less accurately than models, and only slightly more accurately than novices, they seem to have better self-insight about the accuracy of their predictions. Such self-insight is called "calibration." Most people are poorly calibrated, offering erroneous reports of the quality of their predictions, and these reports systematically err in the direction of overconfidence: When they say a class of events are 80% likely, those events occur less than 80% of the time (Lichtenstein, Fischhoff, & Phillips, 1977). There is some evidence that experts are less overconfident than novices. For instance, Levenberg (1975) had subjects look at "kinetic family drawings" to detect whether the children who drew them were normal. The results were, typically, a small victory for training: Psychologists and secretaries got 66% and 61% right, respectively (a coinflip would get half right). Of these cases about which subjects were "positively certain," the psychologists and secretaries got 76% and 59% right, respectively. The psychologists were better calibrated than novices – they used the phrase "positively certain" more cautiously (and appropriately) – but they were still overconfident.

Better calibration of experts has also been found in some other studies (Garb, 1989). Expert calibration is better than novice calibration in bridge (Keren, in press), but not in blackjack (Wagenaar & Keren, 1985). Doctors' judgments of pneumonia and skull fracture are badly calibrated (Christensen-Szalanski & Bushyhead, 1981; DeSmet, Fryback, & Thornbury, 1979). Weather forecasters are extremely well calibrated (Murphy & Winkler, 1977). Experiments with novices showed that training improved calibration, reducing extreme overconfidence in estimating probabilities and numerical quantities (Lichtenstein et al., 1977)

2.6. Summary: expert performance

The depressing conclusion from these studies is that expert judgments in most clinical and medical domains are no more accurate than those of lightly trained novices. (We know of no comparable reviews of other domains, but we suspect that experts are equally unimpressive in most aesthetic, commercial, and physical judgments.) And expert judgments have been worse than those of the simplest statistical models in virtually all domains that have been studied. Experts are sometimes less overconfident than novices, but not always.

3. EXPERT DECISION PROCESSES

The picture of expert performance painted by behavioral decision theorists is unflattering. Why are experts predicting so badly? We know that many experts have special cognitive and memory skills (Chase & Simon, 1973; Ericsson & Polson, 1988; Larkin, McDermott, Simon, & Simon, 1980). Do expert *decision-makers* have similar strategies and skill? If so, why don't they perform better? Three kinds of evidence help answer these questions: process analyses of expert judgments, indirect analyses using regression models, and laboratory studies in which subjects become "artificial experts" in a simple domain.

3.1. Direct evidence: process analyses of experts

The rules and cues experts use can be discovered by using process tracing techniques – protocol analysis and monitoring of information acquisition. Such studies have yielded consistent conclusions across a diverse set of domains.

Search is contingent. If people think like a regression model, weighting cues and adding them, then cue search will be simple – the same variables will be examined, in the same sequence, in every case. Novices behave that way. But experts have a more active pattern of contingent search: Subsets of variables are considered in each case, in different sequences. Differences between novice and expert searches have been found in studies of financial analysts

(Bouman, 1980; E. Johnson, 1988), auditors (Bedard & Mock, 1989), graduate admissions (E. Johnson, 1980), neurologists (Kleinmuntz, 1968), and physicians (Elstein, Shulman, & Sprafka, 1978; P. Johnson, Hassebrock, Duran, & Moller, 1982).

Experts search less. A common finding in studies of expert cognition is that information processing is less costly for experts than for novices. For example, expert waiters (Ericsson & Chase, 1981) and chess players (Chase & Simon, 1973) have exceptional memory skills. Their memory allows more efficient encoding of task-specific information; if they wanted to, experts could search and sit cheaply through more information. But empirical studies show that experts use *less* information than novices, rather than more, in auditing (Bedard, 1989; Bedard & Mock, 1989), financial analysis (Bouman, 1980; E. Johnson, 1988), and product choice (Bettman & Park, 1980; Brucks, 1985; E. Johnson & Russo, 1984).

Experts use more knowledge. Experts often search contingently, for limited sets of variables, because they know a great deal about their domains (Bouman, 1980; Elstein et al., 1978; Libby & Frederick, 1989). Experts perform a kind of diagnostic reasoning, matching the cues in a specific case to prototypes in a casual brand of hypothesis testing. Search is contingent because different sets of cues are required for each hypothesis test. Search is limited because only a small set of cues are relevant to a particular hypothesis.

3.2. Indirect evidence: dissecting residuals

The linear regression models described in section 2 provide a simple way to partition expert judgment into components. The bootstrapped judgment is a linear combination of observed cues; the residual is everything else. By dissecting the residual statistically, we can learn how the decision process experts use deviates from the simple linear combination of cues. It deviates in three ways.

Experts often use configural choice rules. In configural rules, the impact of one variable depends on the values of other variables. An example is found in clinical lore on interpretation of the MMPI. Both formal instruction and verbal protocols of experienced clinicians give rules that note the state of more than one variable. A nice example is given by an early rule-based system constructed by Kleinmuntz (1968) using clinicians' verbal protocols. Many of the rules in the system reflect such configural reasoning: "Call maladjusted if $P_a \geq 70$ unless $M_t \leq 6$, and $K \geq 65$." Because linear regression models weight each cue independently, configural rules will not be captured by the linear form, and the effects of configural judgment will be reflected in the regression residual.

Experts use "broken-leg cues." Cues that are rare but highly diagnostic often are called broken-leg cues, from an example cited by Meehl (1954; pp. 24–

25): A clinician is trying to predict whether or not Professor A will go to the movies on a given night. A regression model predicts that the professor will go, but the clinician knows that the professor recently broke his leg. The cue "broken leg" probably will get no weight in a regression model of past cases, because broken legs are rare.[9] But the clinician can confidently predict that the professor will not go to the movies. The clinician's recognition of the broken-leg cue, which is missing from the regression model, will be captured by the residual. Note that while the frequency of any one broken-leg cue is rare, in "the mass of cases, there may be *many (different) rare kinds of factors*" (Meehl, 1954, p. 25).

Note how the use of configural rules and broken-leg cues is consistent with the process data described in section 3. To use configural rules, experts must search for different sets of cues in different sequences. Experts also can use their knowledge about cue diagnosticity to focus on a limited number of highly diagnostic broken-leg cues. For example, in E. Johnson's (1988) study of financial analysts, experts were much more accurate than novices because they could interpret the impact of news events similar to broken-leg cues.

Experts weight cues inconsistently and make errors in combining them. When experts do combine cues linearly, any inconsistencies in weighting cues, and errors in adding them, will be reflected in the regression residual. Thus, if experts use configural rules and broken-leg cues, their effects will be contained in the residuals of a linear bootstrapping model. The residuals also contain inconsistencies and error. By comparing residual variance and test–retest reliability, Camerer (1981b) estimated that only about 40% of the variance in residuals was error,[10] and 60% was systematic use of configural rules and broken-leg cues. (Those fractions were remarkably consistent across different studies.) The empirical correlation between residuals and outcomes, r_z, however, averaged only about .05 (Camerer, 1981a) over a wider range of studies. Experts are using configural rules and broken-leg cues systematically, but they are not highly correlated with outcomes. Of course, there may be some domains in which residuals are more valid.[11]

3.3. Artificial experts

A final kind of process evidence comes from "artificial experts," subjects who spend much time in an experimental environment trying to induce accurate judgmental rules. A lot of this research belongs to the tradition

[9] Unless a broken leg has occurred in the sample used to derive regression weights, the cue "broken leg" will not vary and will get no regression weight.

[10] These data correct the presumption in the early bootstrapping literature (e.g., Dawes, 1971; Goldberg, 1970) that residuals were entirely human error.

[11] A recent study with sales forecasters showed a higher r_z, around .2 (Blattberg & Hoch, 1990). Even though their residuals were quite accurate, the best forecasters only did about as well as the linear model. In a choice between models and experts, models will win, but a mechanical combination of the two is better still: Adding bootstrapping residuals to an actuarial model increased predictive accuracy by about 10%.

of multiple-cue probability learning (MCPL) experiments that stretches back decades, with the pessimistic conclusion that rule induction is difficult, particularly when outcomes have random error. We shall give three more recent examples that combine process analysis with a rule induction task.

Several studies have used protocol analysis to determine *what* it is that artificial experts have learned. Perhaps the most ambitious attempts to study extended learning in complex environments were Klayman's studies of cue discovery (Klayman, 1988; Klayman & Ha, 1985): Subjects looked at a complex computer display consisting of geometric shapes that affected the distance traveled by ray traces from one point on the display to another. The true rule for travel distance was determined by a complex linear model consisting of seven factors that varied in salience in the display. None of Klayman's subjects induced the correct rule over 14 half-hour sessions, but their performances improved steadily. Some improvement came from discovering correct cues (subjects correctly identified only 2.83 of 7 cues, on average). Subjects who systematically experimented, by varying one cue and holding others fixed, learned faster and better than others. Because the cues varied greatly in how much they affected distance, it was important to weight them differently, but more than four-fifths of the rules stated by subjects did not contain any numerical elements (such as weights) at all. In sum, cue discovery played a clear role in developing expertise in this task, but learning about the relative importance of cues did not.

In a study by Meyer (1987), subjects learned which attributes of a hypothetical metal alloy led to increases in its hardness. As in Klayman's study, subjects continued to learn rules over a long period of time. The true rule for hardness (which was controlled by the experimenter) was linear, but most subjects induced configural rules. Subjects made only fairly accurate predictions, because the true linear rule could be mimicked by nonlinear rules. Learning (better performance) consisted of adding more elaborate and baroque configural rules, rather than inducing the true linear relationships.

In a study by Camerer (1981b), subjects tried to predict simulated wheat-price changes that depended on two variables and a large interaction between them (i.e., the true rule was configural). Subjects did learn to use the interaction in their judgments, but with so much error that a linear bootstrapping model that omitted the interaction was more accurate. Similarly, in E. Johnson's (1988) financial-analyst study, even though expert analysts used highly diagnostic news events, their judgments were inferior to those of a simple linear model.

3.4. Summary: expert decision processes

Studies of decision processes indicate that expert decision makers are like experts in other domains: They know more and use their knowledge to guide search for small subsets of information, which differ with each case. Residuals from bootstrapping models and learning experiments also show that

experts use configural rules and cues not captured by linear models (but these are not always predictive). The process evidence indicates that experts know more, but what they know does not enable them to outpredict simple statistical rules. Why not?

4. RECONCILING THE PERFORMANCE AND PROCESS VIEWS OF EXPERTISE

One explanation for the process–performance paradox is that prediction is only one task that experts must perform; they may do better on other tasks. Later we shall consider this explanation further. Another explanation is that experts are quick to develop configural rules that often are inaccurate, but they keep these rules or switch to equally poor ones. (The same may be true of broken-leg cues.) This argument raises three questions, which we address in turn: Why do experts develop configural rules? Why are configural rules often inaccurate? Why do inaccurate configural rules persist?

4.1. Why do experts develop configural rules?

Configural rules are easier. Consider two common classes of configural rules, conjunctive (hire Hope for the faculty if she has glowing letters of recommendation, good grades, *and* an interesting thesis) and disjunctive (draft Michael for the basketball team if he can play guard *or* forward *or* center extremely well). Configural rules are easy because they bypass the need to trade off different cues (Are recommendations better predictors than grades?), avoiding the cumbersome weighting and combination of information. Therefore, configural rules take much less effort than optimal rules and can yield nearly optimal choices (E. Johnson & Payne, 1985).[12]

Besides avoiding difficult trade-offs, configural rules require only a simple categorization of cue values. With conjunctive and disjunctive rules, one need only know whether or not a cue is above a cutoff; attention can be allocated economically to categorize the values of many cues crudely, rather than categorizing only one or two cues precisely.

Prior theory often suggests configural rules. In his study of wheat prices, Camerer (1981b) found that subjects could learn of the existence of a large configural interaction only when cue labels suggested the interaction a priori. Similarly, cue labels may cause subjects to learn configural rules where they are inappropriate, as in Meyer's (1987) study of alloy hardness. These prior beliefs about cue–outcome correlations often will be influenced by the "representativeness" (Tversky & Kahneman, 1982) of cues to outcomes; the representativeness heuristic will sometimes cause errors.

[12] Configural rules are especially useful for narrowing a large set of choices to a subset of candidates for further consideration.

Besides their cognitive ease and prior suggestion, complex configural rules are easy to learn because it is easy to weave a causal narrative around a configural theory. These coherent narratives cement a dependence between variables that is easy to express but may overweight these "causal" cues, at the cost of ignoring others. Linear combinations yield no such coherence. Meehl (1954) provides the following example from clinicial psychology, describing the case of a woman who was ambivalent toward her husband. One night the woman came home from a movie alone. Then:

> Entering the bedroom, she was terrified to see, for a fraction of a second, a large black bird ("a raven, I guess") perched on her pillow next to her husband's head. . . . She recalls "vaguely, some poem we read in high school." (p. 39)

Meehl hypothesized that the woman's vision was a fantasy, based on the poem "The Raven" by Edgar Allen Poe: "The [woman's] fantasy is that like Poe's Lenore, she will die or at least go away and leave him [the husband] alone." Meehl was using a configural rule that gave more weight to the raven vision because the woman knew the Poe poem. A linear rule, simply weighting the dummy variables "raven" and "knowledge of Poe," yields a narrative that is much clumsier than Meehl's compelling analysis. Yet such a model might well pay attention to other factors, such as the woman's age, education, and so forth, which might also help explain her ambivalence.

Configural rules can emerge naturally from trying to explain past cases. People learn by trying to fit increasingly sophisticated general rules to previous cases (Brehmer, 1980; Meyer, 1987). Complicated configural rules offer plenty of explanatory flexibility. For example, a 6-variable model permits 15 two-way interactions, and a 10-variable model allows 45 interactions.[13] In sports, for instance, statistics are so plentiful and refined that it is easy to construct subtle "configuralities" when global rules fail. Bucky Dent was an average New York Yankee infielder, except in the World Series, where he played "above his head," hitting much better than predicted by his overall average. (The variable "Dent" was not highly predictive of success, but adding the interaction "Dent" × "Series" was.)[14] Because people are reluctant to accept the possibility of random error (Einhorn, 1986), increasingly complicated configural explanations are born.

Inventing special cases is an important mechanism for learning in more

[13] A linear model with k cues has only k degrees of freedom, but the k variables offer $k(k-1)/2$ multiplicative two-variable interactions (and lots of higher-order interactions).

[14] We cannot determine whether Dent was truly better in the World Series or just lucky in a limited number of Series appearances. Yet his success in "big games" obviously influenced the Yankees' owner, George Steinbrenner (who has not otherwise distinguished himself as an expert decision-maker). He named Dent manager of the Yankees shortly after this conference was held, citing his ability as a player "to come through when it mattered." Dent was later fired 49 games into the season (18 wins, 31 losses), and the Yankees had the worst record in Major League baseball at the time.

deterministic environments, where it can be quite effective. The tendency of decision-makers to build special-case rules mirrors more adaptive processes of induction (e.g., Holland, Holyoak, Nisbett, & Thagard, 1986, chapter 3, esp. pp. 88–89) that can lead to increased accuracy. As Holland and associates pointed out, however, the validity of these mechanisms rests on the ability to check each specialization on many cases. In noisy domains like the ones we are discussing, there are few replications. It was unlikely, for example, that Dent would appear in many World Series, and even if he did, other "unique" circumstances (opposing pitching, injuries, etc.) could always yield further "explanatory" factors.

In sum, configural rules are appealing because they are easy to use, have plausible causal explanations, and offer many degrees of freedom to fit data. Despite these advantages, configural rules may have a downfall, as detailed in the next section.

4.2. Why are configural rules often inaccurate?

One reason configural rules may be inaccurate is that whereas they are induced under specific and often rare conditions, they may well be applied to a larger set of cases. Often, people induce such rules from observation, they will be overgeneralizing from a small sample (expecting the sample to be more "representative" of a population that it is – Tversky & Kahneman, 1982). This is illustrated by a verbal protocol recorded by a physician who was chair of a hospital's admissions committee for house staff, interns, and residents. Seeing an applicant from Wayne State who had very high board scores, the doctor recalled a promising applicant from the same school who had perfect board scores. Unfortunately, after being admitted, the prior aspirant had done poorly and left the program. The physician recalled this case and applied it to the new one: "We have to be quite careful with people from Wayne State with very high board scores. . . . We have had problems in the past."

Configural rules may also be wrong because the implicit theories that underlie them are wrong. A large literature on "illusory correlation" contains many examples of variables that are thought to be correlated with outcomes (because they are similar) but are not. For example, most clinicians and novices think that people who see male features or androgynous figures in Rorschach inkblots are more likely to be homosexual. They are not (Chapman & Chapman, 1967, 1969). A successful portfolio manager we know refused to buy stock in firms run by overweight CEOs, believing that control of one's weight and control of a firm are correlated. Because variables that are only illusorily correlated with outcomes are likely to be used by both novices and experts, the small novice–expert difference suggests that illusory correlations may be common.

Configural rules are also likely to be unrobust to small errors, or "brittle."[15]

[15] Although the robustness of linear models is well established, we know of no analogous work on the *un*robustness of configural rules.

Linear models are extremely robust; they fit nonlinear data remarkably well (Yntema & Torgerson, 1961). That is why omitting a configural interaction from a bootstrapping model does not greatly reduce the accuracy of the model.[16] In contrast, we suspect that small errors in measurement may have great impacts on configural rules. For example, the conjunctive rule "require good grades *and* test scores" will lead to mistakes if a test score is not a predictor of success or if the cutoff for "good grades" is wrong; the linear rule that weights grades and scores and combines them is less vulnerable to either error.

4.3. Why do inaccurate configural rules persist?

One of the main lessons of decision research is that feedback is crucial for learning. Inaccurate configural rules may persist because experts who get slow, infrequent, or unclear feedback will not learn that their rules are wrong. When feedback must be sought, inaccurate rules may persist because people tend to search instinctively for evidence that will confirm prior theories (Klayman & Ha, 1985). Even when feedback is naturally provided, rather than sought, confirming evidence is more retrievable or "available" than disconfirming evidence (Tversky & Kahneman, 1973). The disproportionate search and recall of confirming instances will sustain experts' faith in inaccurate configural rules. Even when evidence does disconfirm a particular rule, we suspect that the natural tendencies to construct such rules (catalogued earlier) will cause experts to refine their rules rather than discard them.

4.4. Nonpredictive functions of expertise

The thinking of experts is rich with subtle distinctions, novel categories, and complicated configural rules for making predictions. We have given several reasons why such categories and rules might arise, and persist even if they are inaccurate. Our arguments provide one possible explanation why knowledgeable experts, paradoxically, are no better at making predictions than novices and simple models.

Another explanation is that the knowledge that experts acquire as they learn may not be useful for making better predictions about important long-range outcomes, but it may be useful for other purposes. Experts are indispensable for measuring variables (Sawyer, 1966) and discovering new ones (E. Johnson, 1988).

Furthermore, as experts learn, they may be able to make more kinds of predictions, even if they are no more accurate; we speculate that they mistake their increasing fertility for increasing accuracy. Taxi drivers know lots of alternative routes when they see traffic on the Schuylkill Expressway (cf.

[16] Linear models are robust to nonlinearities provided the relationship between each predictor and outcome has the same direction for any values of the other predictors (although the relationship's magnitude will vary). This property is sometimes called "conditional monotonicity."

Chase, 1983), and they probably can predict their speeds on those alternative routes better than a novice can. But can the experts predict whether there will be heavy traffic on the expressway better than a statistical model can (using time of day, day of week, and weather, for example)? We doubt it.

There are also many social benefits of expertise that people can provide better than models can. Models can make occasional large mistakes that experts, having common sense, would know to avoid (Shanteau, 1988).[17] Experts can explain themselves better, and people usually feel that an expert's intuitive judgments are fairer than those of a model (cf. Dawes, 1971).

Some of these attitudes toward experts stem from the myth that experts are accurate predictors, or the hope that an expert will never err.[18] Many of these social benefits should disappear with time, if people learn that models are better; until then, experts have an advantage. (Large corporations have learned: They use models in scoring credit risks, adjusting insurance claims, and other activities where decisions are routine and cost savings are large. Consumers do think that such rules are unfair, but the cost savings overwhelm their objections.)

5. IMPLICATIONS FOR UNDERSTANDING EXPERT DECISION MAKING

Our review produces a consistent, if depressing, picture of expert decision-makers. They are successful at generating hypotheses and inducing complex decision rules. The result is a more efficient search of the available information directed by goals and aided by the experts' superior store of knowledge. Unfortunately, their knowledge and rules have little impact on experts' performance. Sometimes experts are more accurate than novices (though not always), but they are rarely better than simple statistical models.

An inescapable conclusion of this research is that experts do some things well and others poorly. Sawyer (1966) found that expert measurement of cues, and statistical combination of them, worked better than expert combination or statistical measurement. Techniques that combine experts' judgments about configural and broken-leg cues with actuarial models might improve performance especially well (Blattberg & Hoch, 1990; E. Johnson, 1988).

Of course, expert performance relative to models depends critically on the

[17] This possibility has been stressed by Ken Hammond in discussions of analytical versus intuitive judgment (e.g., Hammond, Hamm, Grassia, & Pearson, 1987). For example, most of the unorthodox moves generated by the leading backgammon computer program (which beat a world champion in 1979) are stupid mistakes an expert would catch; a few are brilliant moves that might not occur to an expert.

[18] A model necessarily errs, by fixing regression coefficients and ignoring many variables. It "accepts error to make less error" (Einhorn, 1986). An expert, by changing regression coefficients and selecting variables, conceivably could be right every time. This difference is made dramatic by a medical example. A statistician developed a simple linear model to make routine diagnoses. Its features were printed on a card doctors could carry around; the card showed several cues and how to add them. Doctors wouldn't use it because they couldn't defend it in the inevitable lawsuits that would result after the model would have made a mistake.

task and the importance of configural and broken-leg cues. There may be tasks in which experts beat models, but it is hard to think of examples. In pricing antiques, classic cars, or unusual real estate (e.g., houses over $5 million), there may be many broken-leg cues that give experts an advantage, but a model including the expert-rated cue "special features" may also do well.

Tasks involving pattern recognition, like judging the prospective taste of gourmet recipes or the beauty of faces or paintings, seem to involve many configural rules that favor experts. But if one adds expert-rated cues like "consistency" (in recipes) or "symmetry" (in faces) to linear models, the experts' configural edge may disappear.

Another class of highly configural tasks includes those in which variable weights change across subsamples or stages. For instance, one should play the beginning and end of a backgammon or chess game differently. A model that picks moves by evaluating position features, weighting them with fixed weights, and combining them linearly will lose to an expert who implicitly changes weights. But a model that could shift weights during the game could possibly beat an expert, and one did: Berliner's (1980) backgammon program beat the 1979 world champion.

There is an important need to provide clearer boundaries for this dismal picture of expert judgment. To what extent, we ask ourselves, does the picture provided by this review apply to the other domains discussed in this volume? Providing a crisp answer to this question is difficult, because few of these domains provide explicit comparisons between experts and linear models. Without such a set of comparisons, identifying domains in which experts will do well is speculation.

We have already suggested that some domains are inherently richer in broken-leg and configural cues. The presence of these cues provides the opportunity for better performance but does not necessarily guarantee it. In addition, the presence of feedback and the lack of noise have been suggested as important variables in determining the performances of both experts and expert systems (Carroll, 1987). Finally, Shanteau (1988) has suggested that "good" experts are those in whom the underlying body of knowledge is more developed, providing examples such as soil and livestock judgment.

6. IMPLICATIONS FOR THE STUDY OF EXPERTISE

Expertise should be identified by comparison to some standard of performance. Random and novice performances make for natural comparisons. The linear-model literature suggests that simple statistical models provide another, demanding comparison.

The results from studies of expert decision making have had surprisingly little effect on the study of expertise, even in related tasks. For instance, simple linear models do quite well in medical-judgment tasks such as the hypothetical task discussed at the beginning of this chapter. Yet most of the

work in aiding diagnosis has been aimed at developing expert systems that can mimic human expert performance, not exceed or improve upon it.

Expert systems may predict less accurately than simple models because the systems are *too much* like experts. The main lesson from the regression-model literature is that large numbers of configural rules, which knowledge engineers take as evidence of expertise, do not necessarily make good predictions; simple linear combinations of variables (measured by experts) are better in many tasks.

A somewhat ironic contrast between rule-based systems and linear models has occurred in recent developments in connectionist models. Whereas these models generally represent a relatively low level of cognitive activity, there are some marked similarities to the noncognitive "paramorphic" regression models we have discussed. In many realizations, a connectionist network is a set of units with associated weights that specify constraints on how the units combine the input received. The network generates weights that will maximize the goodness of fit of the system to the outcomes it observes in training (Rumelhart, McClelland, & PDP Research Group, 1986).

In a single-layer system, each unit receives its input directly from the environment. Thus, these systems appear almost isomorphic to simple regressions, producing a model that takes environmental cues and combines them, in a linear fashion, to provide the best fit to the outcomes. Much like regressions, we would expect simple, single-layer networks to make surprisingly good predictions under uncertainty (Jordan, 1986; Rumelhart et al., 1986).

More complex, multilayer systems allow for the incorporation of patterns of cues, which resemble the configural cues reported by experts. Like human experts, we suspect that such hidden units in these more complex systems will not add much to predictive validity in many of the domains we have discussed. The parallel between regression models and connectionist networks is provocative and represents an opportunity for bringing together two quite divergent paradigms.

Finally, we note that this chapter stands in strong contrast to the chapters that surround it: Our experts, while sharing many signs of superior expert processing demonstrated in other domains, do not show superior performance. The contrast suggests some closing notes. First, the history of the study of expert decision making raises concerns about how experts are to be identified. Being revered as an expert practitioner is not enough. Care should be given to assessing actual performance. Second, the case study of decision making may say something about the development of expertise in general and the degree to which task characteristics promote or prevent the development of superior performance. Experts fail when their cognitive abilities are badly matched to environmental demands.

In this chapter we have tried to isolate the characteristics of decision tasks that (1) generate such poor performance, (2) allow experts to believe that they are doing well, and (3) allow us to believe in them. We hope that the contrast between these conditions and those provided by other domains may

contribute to a broader, more informed view of expertise, accounting for experts' failures as well as their successes.

ACKNOWLEDGMENTS

The authors contributed equally; the order of authors' name is purely alphabetical. We thank Helmut Jungermann, as well as Anders Ericsson, Jaqui Smith, and the other participants at the Study of Expertise conference in Berlin, 25–28 June 1989, at the Max Planck Institute for Human Development and Education, for many helpful comments. Preparation of this chapter was supported by a grant from the Office of Naval Research and by NSF grant SES 88-09299.

REFERENCES

Bedard, J. (1989). Expertise in auditing: Myth or reality? *Accounting, Organizations and Society, 14,* 113–131.

Bedard, J., & Mock, T. J. (1989). *Expert and novice problem-solving behavior in audit planning: An experimental study.* Unpublished paper, University of Southern California.

Berliner, H. J. (1980). Backgammon computer program beats world champion. *Artificial Intelligence, 14,* 205–220.

Bettman, J. B., & Park C. W. (1980). Effects of prior knowledge, exposure and phrase of choice process on consumer decision processes. *Journal of Consumer Research, 17,* 234–248.

Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% database + 50% manager. *Management Science, 36,* 887–899.

Bouman, M. J. (1980). Application of information-processing and decision-making research, I. In G. R. Ungson & D. N. Braunstein (Eds.), *Decision making: An interdisciplinary inquiry* (pp. 129–167). Boston: Kent Publishing.

Bowman, E. H. (1963). Consistency and optimality in management decision making. *Management Science, 10,* 310–321.

Brehmer, B. (1980). In one word: Not from experience. *Acta Psychologica, 45,* 223–241.

Brucks, M. (1985). The effects of product class knowledge on information search behavior. *Journal of Consumer Research, 12,* 1–16.

Brunswik, E. (1952). *The conceptual framework of psychology.* University of Chicago Press.

Camerer, C. F. (1981a). The validity and utility of expert judgment. Unpublished Ph.D. dissertation, Center for Decision Research, University of Chicago Graduate School of Business.

Camerer, C. F. (1981b). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance, 27,* 411–422.

Carroll, B. (1987). Expert systems for clinical diagnosis: Are they worth the effort? *Behavioral Science, 32,* 274–292.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 73,* 193–204.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an otstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 46,* 271–280.

Chase, W. G. (1983). Spatial representations of taxi drivers. In D. R. Rogers & J. H Sloboda (Eds.), *Acquisition of symbolic skills* (pp. 391–405). New York: Plenum.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4,* 55–81.

Christensen-Szalanski, J. J. J., & Bushyhead, J. B. (1981). Physicians' use of probabilistic information in a real clinical setting. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 928–935.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26,* 180–188.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 97.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674.

DeSmet, A. A., Fryback, D. G., & Thornbury, J. R. (1979). A second look at the utility of radiographic skull examination for trauma. *American Journal of Radiology, 132,* 95–99.

Einhorn, H. E. (1974). Expert judgment: Some necessary conditions and an example. *Journal of Applied Psychology, 59,* 562–571.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organization Behavior and Human Performance, 7,* 86–106.

Einhorn, H. J. (1986). Accepting error to make less error. *Journal of Personality Assessment, 50,* 387–395.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemas for decision making. *Organization Behavior and Human Performance, 13,* 171–192.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process tracing models of judgment. *Psychological Review, 86,* 465–485.

Elstein, A. S., Shulman, A. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Ericsson, K. A., & Chase, W. G. (1981). Exceptional memory. *American Scientist, 70*(6), 607–615.

Ericsson, K. A., & Polson, P. G. (1988). An experimental analysis of the mechanisms of a memory skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 305–316.

Ericsson, K. A., & Simon, H. A. (1987). Verbal reports as data. *Psychological Review, 87,* 215–251.

Garb, H. N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105,* 387–396.

Goldberg, L. R. (1959). The effectiveness of clinicians' judgments: The diagnosis of organic brain damage from the Bender-Gestalt test. *Journal of Consulting Psychology, 23,* 25–33.

Goldberg, L. R. (1968). Simple models or simple processes? *American Psychologist, 23,* 483–496.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73,* 422–432.

Gustafson, J. E. (1963). The computer for use in private practice. In *Proceedings of Fifth IBM Medical Symposium,* pp. 101–111. White Plains, NY: IBM Technical Publication Division.

Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review, 62,* 255–262.

Hammond, K. R. (1987). Toward a unified approach to the study of expert judgment. In J. Mumpower, L. Phillips, O. Renn, & V. R. R. Uppuluri (Eds.), *NATO ASI Series F: Computer & Systems Sciences: Vol. 35, Expert judgment and expert systems* (pp. 1–16). Berlin: Springer-Verlag.

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-17,* 753–770.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57,* 116–131.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery.* Cambridge, MA: MIT Press.

Johnson, E. J. (1980). Expertise in admissions judgment. Unpublished doctoral dissertation, Carnegie-Mellon University.

Johnson, E. J. (1988). Expertise and decision under uncertainty: Performance and process. In M. T. H. Chi, R. Glaser & M. J. Farr (Eds.), *The nature of expertise* (pp. 209–228). Hillsdale, NJ: Erlbaum.

Johnson, E. J., & Payne, J. (1985). Effort and accuracy in choice. *Management Science, 31,* 395–414.

Johnson, E. J., & Russo, J. E. (1984). Product familiarity and learning new information. *Journal of Consumer Research, 11,* 542–550.

Johnson, P. E., Hassebrock, F., Duran, A. S., & Moller, J. (1982). Multimethod study of clinical judgment. *Organizational Behavior and Human Performance, 30,* 201–230.

Jordan, M. I. (1986). An introduction to linear algebra in parallel distributed processing. In D. Rumelhart, Rumelhart, J. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations* (pp. 365–422). Cambridge, MA.: MIT Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press.

Keren, G. B. (1987). Facing uncertainty in the game of bridge: A calibration study. *Organizational Behavior and Human Decision Processes, 139,* 98–114.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology:: Learning, Memory, and Cognition, 14,* 317–330.

Klayman, J., & Ha, Y. (1985). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Kleinmuntz, B. (1968). *Formal representation of human judgment.* New York: Wiley.

Kundel, H. L., & LaFollette, P. S. (1972). Visual search patterns and experience with radiological images. *Radiology, 103,* 523–528.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335–1342.

Levenberg, S. B. (1975). Professional training, psychodiagnostic skill, and kinetic family drawings. *Journal of Personality Assessment, 39,* 389–393.

Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance, 16,* 1–12.

Libby, R., & Frederick, D. M. (1989, February). *Expertise and the ability to explain audit findings* (University of Michigan Cognitive Science and Machine Intelligence Laboratory Technical Report No. 21).

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1977). Calibration of probabilities: The state of the art. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs.* Amsterdam: D. Reidel.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Meehl, P. E. (1986). Causes and effects of my disturbing little book. *Journal of Personality Assessment, 50,* 370–375.

Meyer, R. J. (1987). The learning of multiattribute judgment policies. *Journal of Consumer Research, 14,* 155–173.

Murphy, A. H., & Winkler, R. L. (1977). Can weather forecasters formulate reliable probability forecasts of precipitation and temperature? *National Weather Digest, 2,* 2–9.

Philadelphia Inquirer. (1989, August 15). Personality test gets revamped for the '80s, pp. 1-D, 3-D.

Rumelhart, D., McClelland, J., & PDP Research Group (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* Cambridge, MA: MIT Press.

Sarbin, T. R. (1944). The logic of prediction in psychology. *Psychological Review, 51,* 210–228.

Sawyer J. (1966). Measurement and prediction, clinical and statistical. *Psychological Bulletin, 66,* 178–200.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68,* 203–215.

Shortliffe, E. H., Buchanan, B. G., & Feigenbaum, E. A. (1979). Knowledge engineering for medical decision making: A review of computer-based decision aids. *Proceedings of the IEEE, 67,* 1207–1224.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 4,* 207–232.

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 84–98). Cambridge: Cambridge University Press.

Voss, J. F., & Post, T. A. (1988). On the solving of ill-structured problems. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 261–285). Hillsdale, NJ: Erlbaum.

Wagenaar, W. A., & Keren, G. B. (1985). Calibration of probability assessments by professional blackjack dealers, statistical experts, and lay people. *Organizational Behavior and Human Decision Processes, 36,* 406–416.

Wiggens, N., & Kohen, E. S. (1971). Man vs. model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology, 19,* 100–106.

Yntema, D. B., & Torgerson, W. J. (1961). Man–computer cooperation in decision requiring common sense. *IRE Transactions of the Professional Group on Human Factors in Electronics, 2,* 20–26.